The Role of Ontologies for Official Statistics

Monica Scannapieco*
Istat, Rome, Italy – scannapi@istat.it

Raffaella Maria Aracri
Istat, Rome, Italy – raffaella.aracri@istat.it

Roberta Radini
Istat, Rome, Italy – radini@istat.it

Laura Tosco
Istat, Rome, Italy – tosco@istat.it

Abstract

National Statistical Institutes (NSIs) have a long experience in dealing with metadata, but, only recently, they started to use *knowledge representation* approaches to govern and exploit their data asset.

An ontology is a formal description of a domain of interest expressed through a computational language. In order to deal with the complexity of the data and metadata assets of NSIs, ontology-based approaches look very promising.

In this paper, we will illustrate some experiences by the Italian National Institute of Statistics (Istat) on using ontologies for the purpose of data integration on one side and data dissemination on the other side. The shown data integration project is based on the Ontology Based Data Management (OBDM) paradigm, proposed for integrating multiple and heterogeneous data sources. The dissemination experience exploits the Linked Data paradigm, realizing the so-called *Web of Data* that can be seen as a special case of the *Internet of Things*.

Keywords: ontologies, linked open data, data integration

1. Introduction

National Statistical Institutes (NSIs) play an important role as data producers, by publishing Official Statistics in the service of citizens and policy-makers. Statistical production processes are indeed intended to produce “data” as their final output.

A statistical production pipeline mainly consists of: (i) a data collection phase, where both direct data collections, like surveys, and secondary ones, like administrative data acquisitions, are performed; (ii) a data processing and analysis phase where data are corrected, integrated and analyzed and (iii) a data dissemination phase where data are published in accordance to final users’ requirements.

In this paper, we will show the current efforts by Istat to address the (macro) phases (ii) and (iii) by exploiting the powerful instrument of ontologies.

In Section 2, we will address how the emerging Ontology Based Data Management (OBDM) paradigm is being used to design and implement the new Integrated System of Statistical Registers, which will serve as a pillar of Istat’s statistical production.

In Section 3, instead, a specific focus will be done on the use of ontologies to disseminate Istat’s data and metadata.

Finally, Section 4 will draw some conclusions.

2. The Integrated System of Statistical Registers as an Ontology-based Data Integration System

Istat has engaged a modernization programme that includes a significant revision of the statistical production process. The focal point of such an important change is the adoption of a system of integrated statistical registers as a base for all the production surveys; this system will be in the following referred to as the Italian Integrated System of Statistical Registers (ISSR). A system of statistical registers consists of a number of registers that can be linked to each other, specifically:

- Base Statistical Registers: (i) Individuals, Families and cohabitations; (ii) Economic Units; (iii) Places; (iv) Activities.
- Extended Statistical Registers, which extend the information available for a population of a specific Base Statistical Register with other variables.
- Thematic Statistical Registers, which are not bound to a specific population, but rather they have the objective of supporting statistics referred to more than one statistical population.

A first initial effort toward the modeling of the ontology for the Base Statistical Register of Individuals, Families, and Cohabitations, is shown in **Fig. 1**.
The notation used to represent the ontology is the Graphol ([2],[3]) visual language.
The main concepts of the ontology are:

- **Individual**: represents a single person described in terms of gender, date of birth, citizenship, place of birth, educational level and other features;
- **Family**: represents a group of persons tied by marriage, kinship, affinity, adoption, guardianship, or affection, cohabiting and having their usual residence in the same municipality. A family can also be composed by one person.
- **Nuclear Family**: represents a group of persons forming a couple relationship or a parent-child one; namely married or cohabiting couples without children or with never married children, or a single parent with one or more never married children.
- **Cohabitation**: represents a group of persons who, without being tied by marriage, kinship, affinity, spend their lives together for religious reasons, care, assistance, military, prisons and others.

All these concepts are linked by several roles, examples of relevant ones are:

- **Relationship**: represents all kinds of relations linking two Individuals. It is detailed in sub-properties i.e. relativeOf as marriedOf, parentOf (and its inverse son/daughterOf), and affectiveRelationship as cohabitant.
- **Stays**: links an Individual to the concept StayingPlace. This role is detailed in the following ones: usualResidence, livesIn and residence, each describing the different reason why an Individual stays in a place.

The design and implementation of the information architecture of the ISSR relies on the Ontology Based Data Management (OBDM) paradigm [1]. The main reasons underlying this choice are:

- The complexity of the metadata asset (structural metadata asset or intensional data representation) in terms of hugeness and lack of a direct control (several sources are administrative ones that come with their own semantics). The use of ontologies, which permits a formalization and a machine-actionable representation of such metadata, looks promising in order to deal with such a complexity.
- National Statistical Institutes have a long experience in dealing with metadata. However, OBDM has a major difference with respect to approaches typically used within NSIs for designing and implementing metadata management systems, namely: ontologies permit to represent metadata "coupled" with data, so they are not only limited to a "documentation" role but they do permit to "govern" the data integration step by ensuring the quality of integrated data.
- The need for having an integration layer permitting to virtualize data resources and performing "on-the fly" query answering. We think that OBDM can properly answer to such a requirement of the ISSR as an alternative to rigid and materialized traditional data integration approaches like traditional data warehousing.

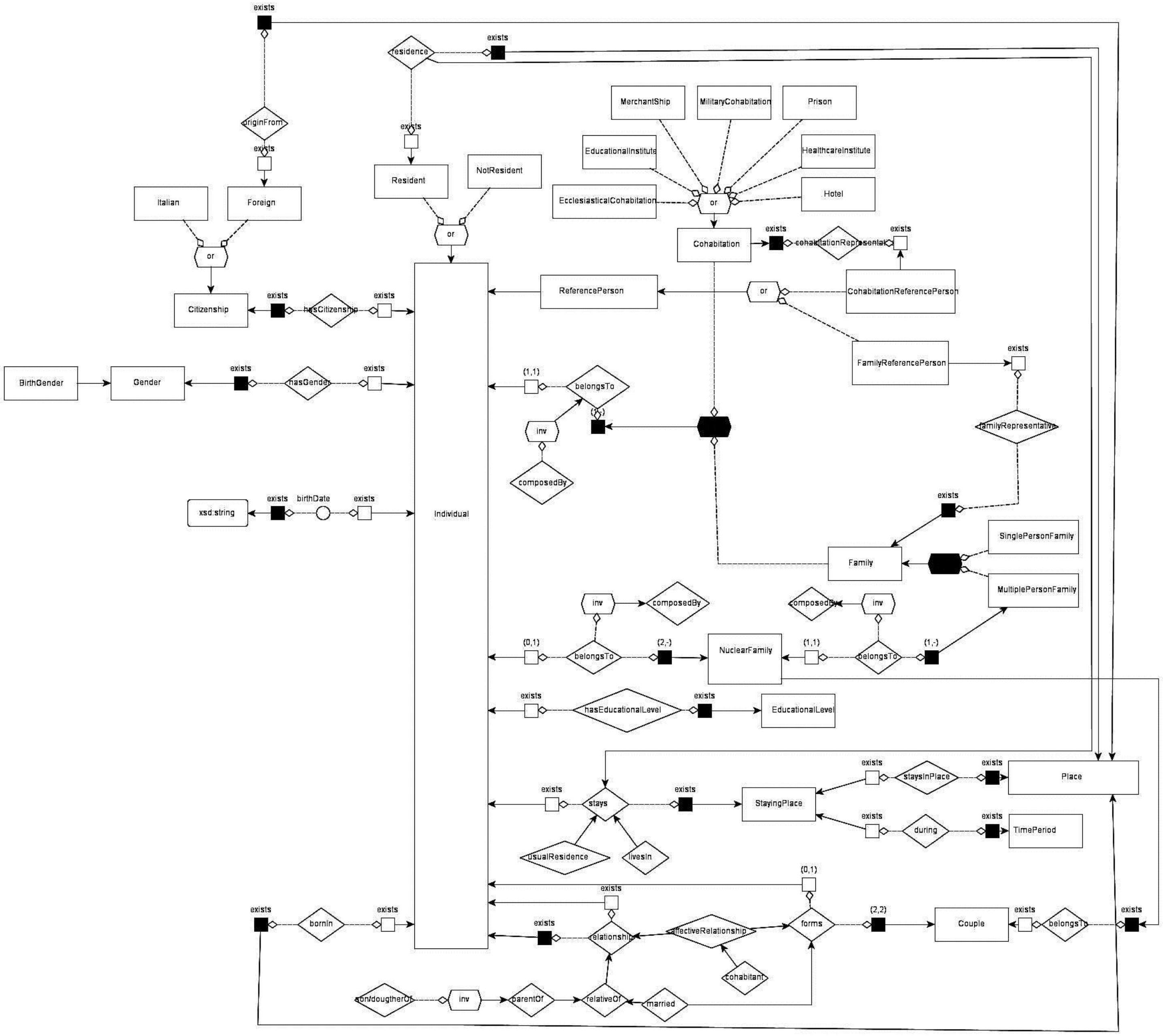


Fig. 1. Ontology for the Base Statistical Register of Individuals, Families and

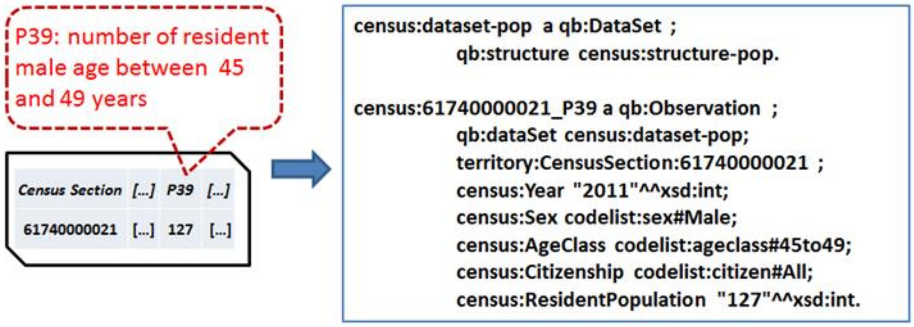
3. The Istat’s Linked Open Data Portal: an Ontology-Based Dissemination Channel

Being “data” the final output of a statistical production process, that is data are the final product that Istat supplies to its end-users, it is very important that released data are semantically enriched. In this respect, the Linked Data paradigm [4], based on ontologies to model data, proved to be extremely suitable as part of the dissemination strategy of Istat.

Istat has published a LOD Portal, available at the URL: datiopen.istat.it. The initial set up was for the purpose of disseminating the 15th Italian Population and Housing Census Data [5]. Data are conceptually modeled through two ontologies expressed according to OWL (Ontology Web Language) [6]:

- the Territorial Ontology that describes the administrative and the geographical organization of the Italian territory. More in detail, the Territorial Ontology describes the administrative organization of the territory, namely: region, province, municipality and geographical-statistical organization of the territory as location, Census section, special areas or special units (like, e.g., abbeys or hospitals).
- the Census Data Ontology that models the actual aggregated Census data. We used the Data Cube Vocabulary [7] to describe the Population and Households census data in terms of measures (e.g., number of residents) and dimensions (e.g., sex or age classes).

Fig. 2 shows an example of an observation expressed through the Data Cube Vocabulary.



Census Section	[...]	P39	[...]
61740000021	[...]	127	[...]

```
census:dataset-pop  a qb:DataSet ;
            qb:structure  census:structure-pop.

census:61740000021_P39 a qb:Observation  ;
            qb:dataSet census:dataset-pop;
            territory:CensusSection:61740000021 ;
            census:Year  "2011"^^xsd:int;
            census:Sex codelist:sex#Male;
            census:AgeClass codelist:ageclass#45to49;
            census:Citizenship  codelist:citizen#All;
            census:ResidentPopulation  "127"^^xsd:int.
```

Fig. 2. Example of a Data Cube observation

Both ontologies make use of meta ontologies as: (i) SKOS [8] for the description of classifications, (ii) ADMS [9] for the description of interoperability assets, and (iii) PROV ontology [10] for the description of the provenance of the data in terms of information about entities, activities, and people involved in the data production process.

The Linked Data paradigm results to be relevant also for metadata dissemination: global uniform naming and addressing are crucial for structural metadata like classifications, code lists, cube dimensions, etc. Istat moved several steps towards the conceptual modeling of classifications [11] to be disseminated through its LOD portal.

Moreover, the statistical community develops metadata standard of good quality, but these standards are “not” represented in formal languages; indeed, such models are mainly described in MS Word documents, XLS files, or UML diagrams that is they are not available internally or for other users in open and machine-actionable formats. Important efforts have been paid to the definition of ontologies for such models (e.g. General Statistical Information Model - GSIM ontology [12], General Statistical Business Process Model - GSBPM ontology [13] and Common Statistical Production Architecture - CSPA ontology [14]). In addition, the need for integrating such models among each other and resolve inconsistencies brought to a specific UNECE project “Implementing ModernStats Standards - Linked Open Metadata” [15]. It is nice to observe how the OWL representation of GSIM, GSBPM and CSPA highlighted some inconsistencies among them (and even within each model) [16].

5. Conclusions

The use of ontologies within an NSI can be relevant for both integration and dissemination purposes. The integration usage is mainly intended for internal statistical users, that can benefit from having:

- Access to integrated data: for instance the "labour" concept has different definitions according to National Accounts, Structural Business Statistics and Labour Force Survey. Ontologies permit that such different definitions can coexist and underlying data can be accessed consistently.
- Reasoning capability: even if some concepts are not "explicitly" linked, reasoning over ontologies allows to "infer" new knowledge (e.g. new relationships). In this way, statistical users can "discover" implicit patterns that can help in understanding data for their analyses.

The dissemination users, mainly external users of Official Statistics, can greatly benefit from the Linked Open Data portal by (i) exploiting services built on machine-to-machine accessible data, and (ii) retrieving data in a transparent way with respect to their physical location (e.g. by accessing simultaneously multiple endpoints).

References

1. M. Lenzerini. Ontology-based data management. In Proc. of the 20th Int. Conf. on Information and Knowledge Management (CIKM 2011), pages 5–6,(2011).
2. Graphol: http://www.dis.uniroma1.it/~graphol
3. Marco Console, Domenico Lembo, Valerio Santarelli, Domenico Fabio Savo: Graphol: Ontology Representation Through Diagrams. In Proc. of the 27th Int. Workshop on Description Logic, 2014.
4. Linked Data: http://linkeddata.org/
5. Raffaella Aracri, Stefano De Francisci, Andrea Pagano, Monica Scannapieco, Laura Tosco, Luca Valentino: Publishing the 15th Italian Population and Housing Census in Linked Open Data. In the Proceedings of the 2nd International Workshop on Semantic Statistics , 2014.
6. Ontology Web Language (OWL): http://www.w3.org/TR/owl-ref/, 10 February 2004
7. Data Cube Vocabulary: http://www.w3.org/TR/2013/CR-vocab-data-cube-20130625/, 25 June 2013
8. Simple Knowledge Organization System (SKOS): http://www.w3.org/TR/2009/REC-skos-reference-20090818/, 18 August 2009
9. Asset Description Metadata Schema (ADMS): https://joinup.ec.europa.eu/asset/adms/home
10. PROV Ontology: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/, 30 April 2013.
11. Giorgia Lodi, Antonio Maccioni, Monica Scannapieco, Mauro Scanu, Laura Tosco: Publishing Official Classification in Linked Open Data. In the Proceedings of the 2nd International Workshop on Semantic Statistics , 2014.
12. M. Scannapieco, L. Tosco, D. Gillman, A. Dreyer, G. Duffes: An OWL Ontology for the Generic Statistical Information Model (GSIM): Design and Implementation. In the Proceedings of the 4th International Workshop on Semantic Statistics, http://ceur-ws.org/Vol-1654/article-03.pdf., 2016.
13. Cotton F., Gillman D.: Modeling the Statistical Process with Linked Metadata. In the Proceedings of the 3th International Workshop on Semantic Statistics, available at http://ceur-ws.org/Vol-1551/article-06.pdf, 2015.
14. A. Dreyer, G. Duffes, F. Cotton: An OWL Ontology for the Common Statistical Production Architecture. In the Proceedings of the 4th International Workshop on Semantic Statistics, http://ceur-ws.org/Vol-1654/article-06.pdf, 2016.
15. IMS – Implementing ModernStats Standard Project http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=122323917
16. Implementing ModernStats Standards Linked Open Metadata Design Guidelines http://www1.unece.org/stat/platform/download/attachments/129172661/HLG-MOS%20-%20IMS%20Design%20Guidelines_Jan2017.docx?version=1&modificationDate=1483969944574&api=v2